Geoparsing Biodiversity Heritage Library Collections: A Preliminary Exploration*

Gretchen R. Stahlman[1][0000−0001−8814−863X] and Carolyn Sheffield[2][0000−0001−9704−7363]

[1] University of Arizona School of Information
[2] Smithsonian Libraries

Abstract. A short pilot study was conducted to provide recommendations on methods and workflows for extracting geographic references from the text of Biodiversity Heritage Library collections and disambiguating these references. An initial survey of the literature was conducted, and a variety of possible techniques and software were subsequently explored for natural language processing, machine learning, document annotation, and map visualization. A test corpus was evaluated, and preliminary findings identify challenges for a full-scale effort towards automated geoparsing, including: varying OCR quality, diversity of the corpus, historical context, and ambiguity of geographic references. The project background, approaches, and preliminary assessment are described here.

Keywords: Biodiversity Heritage Library · geoparsing · biodiversity · text mining · data.

1 Introduction

The short pilot study presented here was conducted for the Biodiversity Heritage Library (BHL), which is headquartered at Smithsonian Libraries. As a project implemented by the Summer 2018 LEADS-4-NDP doctoral fellowship program, the 10-week study focused on exploring methods for automating the identification and disambiguation of geographic names in BHL collections (approximately 55 million scanned pages) and, where possible, translating toponyms to polygons or point locations for visual browsing. By the conclusion of the fellowship project, a survey of the literature and related projects was conducted, and a variety of possible techniques and software were explored for natural language processing, machine learning, document annotation, and map visualization. A test corpus of 50 documents was evaluated, and preliminary findings identified the following challenges for a full-scale effort towards automated geoparsing: varying OCR quality, diversity of the corpus, historical context, and ambiguity of geographic references. The background, approach taken, and preliminary findings are described below.

* Supported by LEADS-4-NDP: http://cci.drexel.edu/mrc/research/leads/

2 Background

As access to curated historical data is increasingly possible through advances in technology and community-driven initiatives, enabling time domain research is a focus for development of resources across disciplines. Scientific literature is potentially a rich source of such information, where extracting data from the text of historical publications can produce insight and augment newer data. For biodiversity and ecological research in particular, Bowker [2] illustrates the importance of data diversity across disciplines for cataloging and studying life itself over vast time periods and geographies. Similarly, Thompson, et al. [15] emphasize that studying complex biological and ecological systems requires techniques for mining historical data to address cross-cutting research questions. A survey on the needs and practices of biodiversity scientists conducted in 2010-11 supports these recommendations [5], with a high percentage of respondents indicating a need for occurrence, distribution, and abundance data, as well as georeferenced collections.

Georeferenced databases are particularly useful for studying the distribution of biodiversity over time. The Global Biodiversity Information Facility (GBIF) [7] supports a growing international database of curated species occurrence data, providing standards and tools for sharing and citation of georeferenced information. As an Associate Participant in GBIF, BHL is a central resource for the research community and embodies a shared mission to make biodiversity data openly accessible and discoverable. BHL also partners with the Encyclopedia of Life (EOL), a rich resource curating descriptive information on all known life on Earth. Both EOL and BHL cross-link taxonomic references across their platforms. BHL collections span 500 years and are contributed by institutions worldwide, consisting of millions of scanned pages with corresponding OCR text files. Currently, BHL offers full text searching, as well as scientific name recognition to display scientific names identified on each page.

Automatically detecting, disambiguating and visualizing place name references in large text datasets is a computing challenge. However, this task also presents a shared opportunity for cutting edge research and development of scalable technology and techniques. For example, using machine learning Weissenbacher, et al. [16] created an automated method for retrieving geospatial metadata from full-text phylogeology literature, successfully implementing toponym detection and disambiguation using dictionary-based and rule-based heuristics, and referring to the GeoNames database of locations for toponym resolution. In contrast to modern scientific literature with generally predictable structural characteristics and topic areas across publications, heterogeneous collections of historical texts are particularly challenging and unsuited for generic named entity recognition algorithms and modern gazetteers alone. To perform toponym resolution on a corpus of Civil War memos, DeLozier, et al. [6] tackled this issue by creating the GeoAnnotate tool for collaborative annotating of locations and coordinates, with likely place names pre-annotated using the StanfordNER package to minimize the workload of annotators. As an example of a solution to the problem of discrepancies between historical and contemporary maps, Cura, et

al. [4] extracted information from historical maps of Paris and matched them to current addresses in modern Paris. For biodiversity, Cardoso, et al. [3] used several existing online geographic databases to create a new open gazetteer with superior recall for adding coordinates to species occurrence records. These projects highlight applicable methods and difficulties associated with attempting to accurately and automatically extract place names from the BHL corpus, which spans not only centuries, but the entire globe.

3 Approach

As a short-term project, the approach undertaken focused on exploratory work including a literature review and stakeholder interviews; testing of select tools and methods; and a comparative assessment of those tools and methods with recommendations for scaling up an approach for a corpus of the size of BHL. Details on each approach are briefly outlined below.

Literature review and exploration of methods. A thorough survey of the literature was initially conducted, and review of the literature and other relevant projects and methods was ongoing throughout the project.

Interviews with stakeholders and researchers. A series of introductory meetings was initiated by BHL with stakeholders and researchers internal and external to Smithsonian. Other formal and informal meetings were held with experts in NLP and biodiversity informatics, as well as with researchers working on similar projects.

Possible methods identified. The BHL collections represent a large full-text dataset. Current computing technology could be used to extract information from the collection on a large scale. Human effort for information extraction and quality control could be crowdsourced by leveraging existing volunteer and citizen science communities.

Annotated test corpus for assessment. For evaluation purposes, a test corpus of 50 documents was selected based on the following criteria: publication in the last 200 years (for better OCR), English language, and concise range of topics.

Demonstration with StanfordNER and Mordecai. Named Entity Recognition (NER) is a natural language processing task, enabled by a variety of available software tools and statistical methods. For this project, two promising resources (StanfordNER and Mordecai Geoparser) were used for demonstration of a possible workflow pipeline leading to visualization of the place names in a BHL document.

4 Preliminary Findings

Through the course of this exploratory project, some NLP challenges particular to the BHL corpus have been identified and are listed here:

OCR quality. OCR is a challenge for NLP tasks across research areas [9, 10], including for scanned biodiversity texts, which often include handwritten field notes [8, 11, 12]. Particularly for older documents, the text will likely require additional human effort to clean errors and/or transcribe the documents prior to processing.

Diversity of the corpus. The diversity of a corpus can affect the performance, and thus the suitability, of machine learning for extracting accurate information [13, 14]. BHL collects a tremendous diversity of texts, ranging from handwritten documents to structured scientific articles in a variety of languages. NER efforts could be clustered according to similarity of document topics, types and languages, while leveraging community input to prioritize areas with higher likelihood of resulting in useful occurrence data.

Historical context. Aside from linguistic variations in historical documents, toponym resolution is particularly challenging for older texts [6]. Cura, et al. [4] refer to historical maps of the city of Paris in order to translate place name references to current addresses a tactic that would not be practical considering the international scope of BHL documents. Furthermore, political boundaries and geographic features (such as rivers and coastlines) also change over time. This creates difficulty in extracting accurate data from the texts and requires contextual insight and customized heuristics.

Ambiguity of geographic references. Some BHL documents contain very structured information such as lists of species and treatments with predictable geographic references in close proximity to species and specimens within the text. However, even within these structured areas, references to places can be ambiguous.

5 Conclusion

The project explored by this pilot study could result in potentially useful biodiversity data if fully implemented. Furthermore, efforts towards a full project could have implications for additional discoveries, such as advances in NLP by developing new training data and models for historical documents. While trying and testing a wide variety of software platforms, this project has also encountered with the ephemeral nature of technology. A number of tools and methods were explored, and the tangible outcomes of the project include an extensive report documenting the insights obtained, as well as plans for a possible community workshop to establish a strategy and resources for moving forward, along with a framework for a 2019 LEADS Fellow to participate and build on this work. Considering that a portion of the BHL corpus can be addressed using existing technologies, areas of the corpus that are of high value to researchers should be prioritized, with input from the research community and other stakeholders.

References

1. Alex B, Burns J (2014) Estimating and rating the quality of optically character recognised text. In: Anonymous Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage ACM, p 97-102.
2. Bowker GC (2000) Biodiversity datadiversity. Soc Stud Sci 30(5):643-683.
3. Cardoso SD, Serique KJ, Amanqui FK et al (2014) A gazetteer for biodiversity data as a linked open data solution. In: Anonymous WETICE Conference (WETICE), 2014 IEEE 23rd International IEEE, p 435-440.
4. Cura R, Dumenieu B, Abadie N et al (2018) Historical collaborative geocoding. ISPRS International Journal of Geo-Information 7(7):262.
5. Davis MLS, Tenopir C, Allard S et al (2014) Facilitating access to biodiversity information: a survey of users needs and practices. Environ Manage 53(3):690-701.
6. DeLozier G, Wing B, Baldridge J et al (2016) Creating a novel geolocation corpus from historical texts. In: Anonymous Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), p 188-198.
7. GBIF: What is GBIF?, https://www.gbif.org/what-is-gbif. Last accessed 1 Sept 2018
8. Heidorn PB, Zhang Q (2013) Label annotation through biodiversity enhanced learning.
9. Klein E, Alex B, Grover C et al (2014) Digging Into Data White Paper: Trading Consequences.
10. Kumar A (2016) A survey on various OCR errors. International Journal of Computer Applications 143(4):8-10.
11. Paul DL, Heidorn PB (2013) Augmenting optical character recognition (OCR) for improved digitization: Strategies to access scientific data in natural history collections.
12. Paul DL, Heidorn PB, Best J et al (2013) Help iDigBio reveal hidden data: iDigBio Augmenting OCR working group needs you-Part II.
13. Shmanina T, Zukerman I, Yepes AJ et al (2013) Impact of corpus diversity and complexity on ner performance. In: Anonymous Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013), p 91-95.
14. Sokolova M, Bobicev V (2018) Corpus Statistics in Text Classification of Online Data. arXiv preprint arXiv:1803.06390.
15. Thompson JN, Reichman O, Morin PJ et al (2001) Frontiers of Ecology: As ecological research enters a new era of collaboration, integration, and technological sophistication, four frontiers seem paramount for understanding how biological and physical processes interact over multiple spatial and temporal scales to shape the earth's biodiversity. AIBS Bulletin 51(1):15-24.
16. Weissenbacher D, Tahsin T, Beard R et al (2015) Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. Bioinformatics 31(12):i348-i356.